DOCUMENT RESUME

ED 137 351 TM 006 152

AUTHOR Hunter, John E.; Schmidt, Frank L.
TITLE Fairness of Selection Tests: A Critical Analysis. Professional Series 76-5.
INSTITUTION Civil Service Commission, Washington, D.C. Personnel Measurement Research and Development Center.
PUB DATE Sep 76
NOTE 44p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Culture Free Tests; *Ethics; *Minority Groups; *Personnel Selection; *Statistical Analysis; *Test Bias

ABSTRACT

The first section of this paper defines three incompatible ethical positions in regard to the fair and unbiased use of psychological tests for selection in minority and majority groups. Also in this section, five statistical definitions of "test fairness" are reviewed and examined critically for technical, logical, and social weaknesses. In the second section of the paper, the various statistical definitions are shown to correlate with specific ethical positions, and the technical, logical, and social problems of each statistical model-ethical position combination are delineated. It is concluded that it is difficult, if not impossible, to predict at the present time which model will ultimately prove most acceptable to the American people. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

Personnel
Research and
Development
Center

Professional Series 76-5

Fairness of Selection Tests: A Critical Analysis

IM006 152

United States
Civil Service Commission
Bureau of Policies and Standards

PS-76-5

FAIRNESS OF SELECTION TESTS: A CRITICAL ANALYSIS

John E. Hunter
Michigan State University

Frank L. Schmidt
Personnel Research and Development Center
U.S. Civil Service Commission

Personnel Research and Development Center
U.S. Civil Service Commission
1900 E Street N.W.
Washington, D.C. 20415
September 1976

FAIRNESS OF SELECTION TESTS: A CRITICAL ANALYSIS

ABSTRACT

The first section of this paper defines three incompatible ethical positions in regard to the fair and unbiased use of psychological tests for selection in minority and majority groups. Also in this section, five statistical definitions of "test fairness" are reviewed and examined critically for technical, logical, and social weaknesses. In the second section of the paper, the various statistical definitions are shown to correlate with specific ethical positions, and the technical, logical, and social problems of each statistical model-ethical position combination are delineated. It is concluded that it is difficult, if not impossible, to predict at the present time which model will ultimately prove most acceptable to the American people.

PREFACE

Those of us concerned with personnel selection and placement cannot but be aware of the ever-increasing involvement of the law and the courts in our professional work. This process has gone so far that questions and issues that only yesterday were regarded by most in the profession as highly specialized and esoteric have become focal points in important, precedent-setting litigation. The problem of defining test fairness given equal test validity coefficients in the relevant applicant subgroups is one such issue. One court decision focusing specifically on this issue is now on record, and others are certain to follow. The purpose of this publication is to provide psychologists, lawyers, and administrators with a thorough yet readable exploration of the issues and problems in this critical area.

CONTENTS

		Page
I.	STATISTICAL AND ETHICAL IMPLICATIONS OF FIVE DEFINITIONS OF "TEST FAIRNESS"	
	John E. Hunter and Frank L. Schmidt	1
	Three Ethical Positions	1
	Unqualified Individualism	1
	Qualified Individualism and the Merit Principle	2
	The Quota Ethic	2
	Five Attempts to Define Test Fairness Statistically	3
	The Cleary Definition	3
	The Thorndike Definition	8
	Darlington's Definition No. 3	10
	Darlington's Definition No. 3 and Cole's Argument	11
	Darlington's Definition No. 4	18
	A fifth Definition of Test Fairness	19
II.	ETHICAL POSITIONS, STATISTICAL DEFINITIONS, AND PROBLEMS	
	Frank L. Schmidt	21
	Unqualified Individualism: Models and Problems	21
	Qualified Individualism: Models and Problems	23
	The Quota Ethic: Models and Problems	25
	Ethical Systems, Statistical Models, Individual Merit, and Social Goals	27
	Figure 1. A case in which the white regression line underpredicts black performance	4
	Figure 2. Regression artifacts produced by unreliability in a Cleary-defined "unbiased" test. A is the common regression line for a perfectly reliable test. B and C are the regression lines for whites and blacks respectively for a test of reliability .50.	6
	Figure 3. Darlington's (1971) method of altering the data to define a "culturally optimal" test.	27
	APPENDIX	29
	FOOTNOTES	33
	REFERENCES	37

I. STATISTICAL AND ETHICAL IMPLICATIONS OF FIVE DEFINITIONS OF "TEST FAIRNESS"

John E. Hunter
Michigan State University

Frank L. Schmidt
Personnel Research and Development Center
U.S. Civil Service Commission

In the last several years there has been a series of papers devoted to the question of the fairness of employment and educational tests to minority groups, (Cleary, 1968; Thorndike, 1971; Darlington, 1971). Although each of these papers came to an ethical conclusion, the basis for that ethical judgment was left unclear. If there were only one ethically defensible position regarding test fairness, then this would pose no problem. But such is not the case. The papers which we shall review have a second common feature. Each writer attempts to establish a definition of test fairness on purely statistical grounds, i.e., on a basis that is independent of the content of test and criterion and which makes no explicit assumption about the causal explanation of the statistical relations found. We will argue that this merely makes the substantive considerations implicit rather than explicit.

In this paper we first describe three distinct ethical positions. We will next examine five statistical definitions of test fairness in detail and show how each is based on one of these ethical positions. Finally, we shall examine the technical, social, and legal advantages and disadvantages of the various ethical positions and statistical definitions.

Three Ethical Positions

Unqualified Individualism

The first ethical position we shall examine, unqualified individualism, defines a fair selection, promotion, or admissions policy as one which uses the best statistical information available - and all of that information - to predict each candidate's future performance and then selects or admits those with the highest predicted performance.

From this point of view, there are two ways in which an institution could act unethically. First it might knowingly fail to use an available more valid predictor, e.g., it might select on the basis of a candidate's appearance rather than his scores on a valid ability test. Secondly, it might deliberately omit a valid predictor that is known to be available, e.g., it might exclude (for trivial reasons) a valid predictor from the regression equation. If race, sex, or ethnic group membership is, in fact, a valid predictor of performance in a given situation, over and above the effects of other measured variables, the unqualified individualist is ethically bound to use such a predictor.

Qualified Individualism and the Merit Principle

This ethical position differs from unqualified individualism in that it specifically forbids the use of illegal or unconstitutional predictors, no matter how valid. If, in a given situation, race is in fact a valid predictor of performance, i.e., the difference between the races on the criterion is greater than would be predicted from the best measures of individual qualifications available, then use of race to predict future job performance is forbidden. Race constitutes an illegal predictor, and its use would be discriminatory. To the unqualified individualist, on the other hand, failure to use race as a predictor would be unethical and discriminatory, since it would result in a less accurate prediction of the future performance of applicants and would "penalize" or underpredict performance of individuals from one of the applicant groups. Unlike the unqualified individualist, the qualified individualist relies solely on measures of ability and motivation to perform the job, e.g., scores on valid aptitude and achievement tests, assessments of past work experiences, etc.

The Quota Ethic

Most corporations and educational institutions are creatures of the state or city in which they function. Thus, it has been argued that they are ethically bound to act in a way which is "politically appropriate" to their location. In particular, in a city whose population is 45 percent black and 55 percent white, any selection procedure which admits any other ratio of blacks and whites is "politically biased" against one group or the other. That is, it is assumed that any politically well defined group has the right to ask and receive its "fair share" of any desirable product or position which is under state control. These fair share

quotas may be based on population percentages or on other factors irrelevant to predicted future performance of selectees (Thorndike, 1971; Darlington, 1971).

Five Attempts To Define Test Fairness Statistically

In this section we will briefly review five attempts to arrive at a statistical criterion for a fair or unbiased test. For ease of presentation, the discussion will be in terms of comparing blacks and whites. However, the reader should bear in mind that other demographic classifications, such as social class or sex, could be substituted with no loss of generality.

The Cleary Definition

Cleary (1968) defined a test to be "unbiased" only if the regression lines for blacks and whites are identical. The reason for this is brought out in Figure 1, which shows a hypothetical case in which the regression line for blacks lies above the line for whites and is parallel to it. Consider a white and a black subject who each have a score of A on the test. If the white regression line were used to predict both criterion scores, then the black applicant would be underpredicted by an amount y, the difference between his expected score making use of the fact he is black and the expected score assigned by the white regression line. Actually in this situation, in order for a white subject to have the same expected performance as a black whose score is A, the white subject must have a score of B.

That is, if the white regression line underpredicts black performance, then a white and black are only truly equal in their expected performance if the white's test score is higher than the black's by an amount related to the amount of underprediction. Similarly, if the white regression line always overpredicts black performance, then a black subject has equal expected performance only if his test score is higher than the corresponding white subject's score by an amount related to the amount of overprediction. If the regression lines for blacks and whites are not equal, then each person will receive a statistically valid predicted criterion score only if separate regression lines are used for the two races. If the two regression lines have exactly the same slope, statistically unbiased prediction could be accomplished by predicting

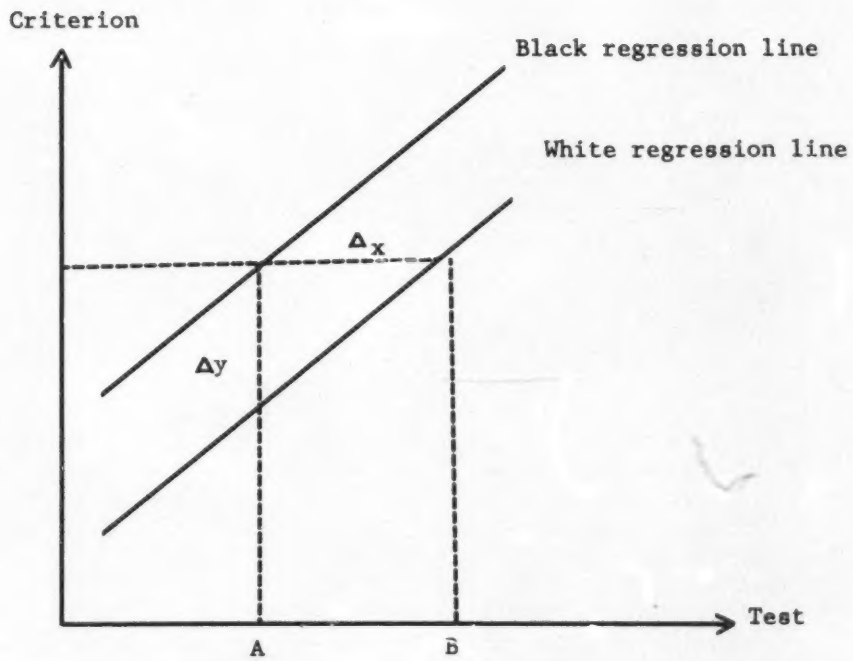


Figure 1. A case in which the white regression line underpredicts black performance.

performance from two separate regression equations or from a multiple regression equation with test score and race as the predictors. If the slopes are not equal, then either separate equations must be used or the multiple regression equation must be expanded by the usual product term for moderator variables. We can therefore view Cleary's definition of an unbiased test as an attempt to rule out disputes between qualified and unqualified individualism. If the predictors available to an institution are "unbiased" in Cleary's sense, then the question of whether or not to use race as a predictor does not arise. But if the predictors are "biased," the recommended use of separate regression lines is clearly equivalent to using race as a predictor of performance. Thus while Cleary (1968) may show a preference for tests that meet the requirements of both unqualified and qualified individualism, her position is, in the final analysis, one of unqualified individualism.

A Cleary-defined "unbiased" test is ethically acceptable under the population-based quota ethic only under very special circumstances. In addition to identical regression lines, blacks and whites must have equal means and equal standard deviations on the test, and this in turn implies equal means and standard deviations on the performance measure. Furthermore, the proportion of black and white applicants must be the same as their proportion in the relevant population. These are conditions that rarely obtain.

Linn and Werts (1971) have pointed out an additional problem for Cleary's definition: The problem of defining fairness when using less than perfectly reliable tests. Suppose that a perfectly reliable measure of ability were in fact an unbiased predictor in Cleary's sense. But since perfect reliability is unattainable in practice, the test actually used will contain a certain amount of error variance. Will the imperfect test be "unbiased" in terms of the regression equations for blacks and whites? If black applicants have a lower mean score than white applicants, then the regression lines for the imperfect test will not be equal. This situation is illustrated in Figure 2. In this figure we see that if an unreliable test is used, then that test produces the double regression line of a "biased" test in which the white regression line over-predicts black performance. That is, by Cleary's definition, the unreliable test is "biased" against whites in favor of blacks. [1,2]

One may be inclined to question whether failure to attain perfect reliability - impossible under any circumstances - should be adequate grounds for labeling a test as biased. But suppose we

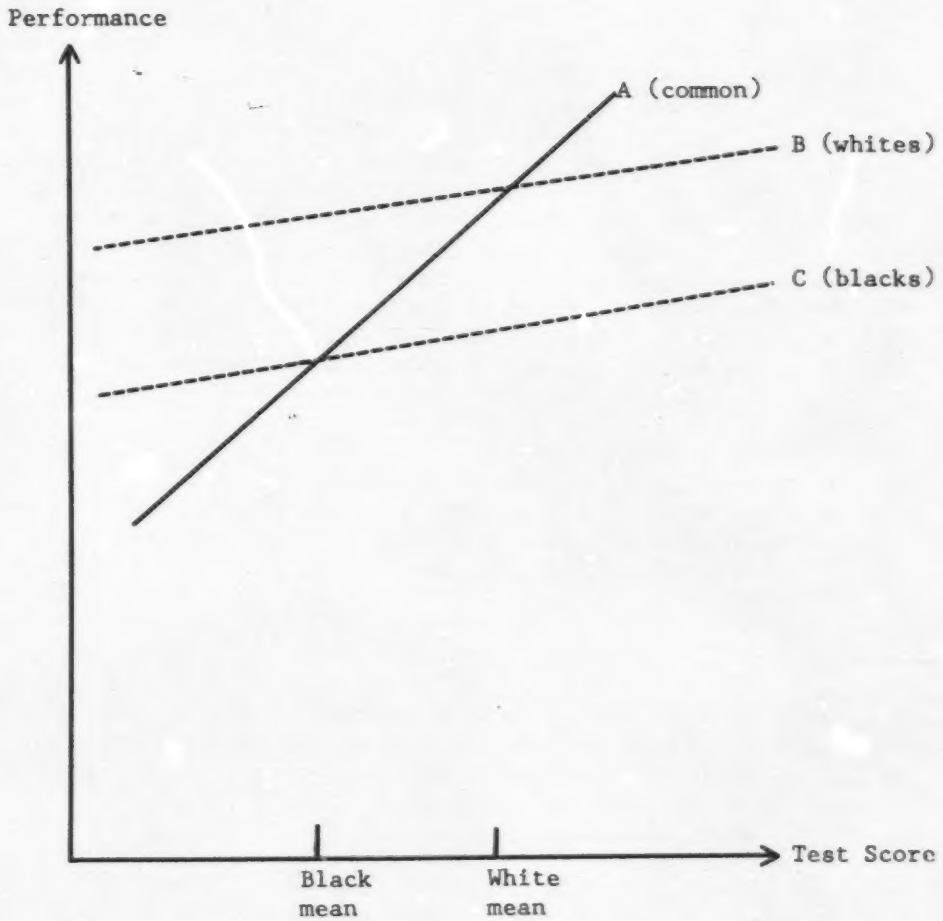


Figure 2. Regression artifacts produced by unreliability in a Cleary-defined "unbiased" test. A is the common regression line for a perfectly reliable test. B and C are the regression lines for whites and blacks respectively for a test of reliability .50.

consider this question from a different viewpoint. Suppose there were only one ethnic group, whites, for example. Assume that Bill has a true ability score of 115 and Jack's true score is 110. If true ability is a valid predictor of performance in this situation, then Bill has the higher expected performance, and if a perfectly reliable test is used, Bill will invariably be admitted ahead of Jack. But suppose that the reliability is only .50. Then the two obtained scores will each vary randomly from their true values and there is some probability that Bill's will be randomly low while Jack's is randomly high, i.e., some probability that Jack will be admitted ahead of Bill. If the standard deviation of the observed ability scores is 15, then the difference between their observed scores has a mean of 5 and a standard deviation of 21. The probability of a negative difference is then .41. Thus the probability that Bill will be admitted ahead of Jack drops from 1.00 to .59. The unreliable test is in fact sharply "biased" against better qualified applicants. It is obvious, however, that this bias is not directly racial or cultural in nature. When there are both blacks and whites in the applicant pool, it takes on the appearance of a racial bias because the proportion of better qualified applicants is higher in the white group. Thus the bias created by random error works against more applicants in the majority group, and on balance the test is "biased" against that group as a whole. But at the individual level, such a test is no more biased against a well qualified white than a well qualified black. The question, then, is whether Cleary's (1968) definition is defective in some sense in labeling this situation as "biased." If so, it may perhaps be desirable to modify the definition to apply only to "bias" beyond that expected on the basis of test reliability alone.[3]

Let us consider the comparison of whites and blacks on the unreliable test in detail. We first remark that it is a fact that on the average, the whites with a given score have a higher mean performance than the blacks who have that same score. Thus the use of a single regression line will, in fact, mean that the whites near the cutoff will be denied admission in favor of blacks who will, on the average, not perform as well. Cleary's (1968) definition would clearly label such a situation "biased." Furthermore, in this situation the partial correlation between race and performance with observed test score held constant is not zero. Thus race makes a contribution to the multiple regression because, with an unreliable test, race is in fact a valid predictor of performance after test score is partialed out. From the viewpoint of unqualified individualism, the failure to use race as a second predictor is unethical. If the test is used with only one regression line, then the predictors

are in fact "biased" against whites as a group. If two regression lines are used, then each person is being considered solely on the basis of expected performance.

For this reason, we feel that this criticism of Cleary's definition is essentially unwarranted. To use an unreliable predictor is to blur valid differences between applicants and an unreliable test is thus, to the extent of the unreliability, biased against applicants or groups of applicants who have high true scores on the predictor. Thus from the point of view of an unqualified individualist, an unreliable test is indeed "biased." On the other hand, a qualified individualist would object to this conclusion. Use of separate regression lines is statistically optimal because the unreliable test does not account for all the real differences on the true scores. But the qualified individualist is ethically prohibited from using race as a predictor and therefore can employ only a single regression equation. He can, however, console himself with the fact that the "bias" in the test is not specifically racial in nature. And he can, of course, attempt to raise the reliability of the test.

The Thorndike Definition

Thorndike (1971) has defined a test as fair if, and only if, subgroup mean differences in standard score form are equal on the test and the criterion. Thorndike noticed that while using two regression lines is often the only ethical procedure from the point of view of unqualified individualism, it need not be required by a specific kind of quota ethic. In particular, if the black regression line is lower, then normally blacks would show a lower mean on both predictor and criterion. Suppose that blacks are one standard deviation lower on both and that validity is .50 for both groups. If we knew the actual criterion scores and set the cut-off at the white mean on the criterion, then fifty percent of the whites and sixteen percent of the blacks would be selected. However, if the predictor score is used with two regression lines, then fifty percent of the whites but only two percent of the blacks will be admitted.[4] Thorndike then argues that this is "unfair" to blacks "as a group." He then recommends that we throw out individualism as a ethical imperative and replace it with a specific kind of quota. The quota that he defines as the "fair share" for each group is the percentage of that group that would have been selected had the criterion itself been used or had the test had perfect validity. In the above situation, for example, Thorndike's definition would consider the selection procedure fair only if 16 percent of the black applicants were selected.

Once Thorndike's definition is shown to be a form of quota setting (c.f. also Schmidt and Hunter, 1974), then the obvious question is "Why his quotas?" After all, the statement "sixteen percent of the blacks can perform at the required level" would not apply to the blacks actually selected and is in that sense irrelevant. In any event, it seems highly unlikely that this method of setting quotas would find support among those adherents of the quota ethic who focus on population proportions as the proper basis of quota determination. Thorndikean quotas will generally be smaller than population-based quotas. On the other hand, Thorndike-determined quotas may have considerable appeal to large numbers of Americans as a seemingly fair compromise between the requirements of qualified individualism and the merit principle, on the one hand, and the social need to upgrade employment levels of minority group members, on the other.

There is another question that must be raised of Thorndike's position: Is it ethically compatible with the use of imperfect selection devices? Assume, for example, that one is using a test score of 50 (X=50, SD=10) as a cutoff and knows from past data that fifty percent of those at or above this test score will be successful. Applicants with test scores of 49 can then correctly state "if we were all admitted, then 49 percent of us would succeed. Therefore according to Thorndike, 49 percent of us should be admitted. Yet we were all denied. Thus, you have been unfair to our group, those people with scores of 49 on the test." Thus, strictly speaking, Thorndike's ethical position precludes the use of any predictor cut-off in selection--no matter how reasonably determined. Instead, from each predictor category one must select that percentage which would fall above the criterion cut-off if the test were perfectly valid. For example, if one wanted to select 50 percent of applicants and the validity were .60, then he would have to take 77 percent of those who lie one standard deviation above the mean, fifty percent of those within one SD of the mean, and 23 percent of those who fall one standard deviation below the mean. And Thorndike's definition could be interpreted, of course, as requiring the use of even smaller intervals of test scores.

There are at least two problems with this procedure. First, one must attempt to explain to applicants with objectively higher qualifications than some selectees why they were not admitted--a rather difficult task and, from the point of view of individualism, an unethical one. Second, the general level of performance will be considerably lower than if the usual cutoff had been used.

In the previous example, the mean performance of the top fifty percent on the predictor would be .48 standard score units, while the mean performance of those selected by the Thorndike ethic would be .29. That is, in this example using Thorndike's quotas has the effect of cutting the utility of the predictor by about 60 percent. (These calculations are shown in the Appendix.)

One possible reply to the criticism that the Thorndike definition leads to a proliferation of subgroups would maintain that it need not be interpreted as requiring application to all definable groups. The definition is to be applied only to "legitimate minority groups," and this would exclude groups defined solely by obtained score on the predictor. If agreement could be reached that, for example, blacks, Chicanos, and Indians are the only recognized minority groups, the definition might be workable. But such an agreement is highly unlikely. On what grounds could we fairly exclude Polish, Italian, and Greek Americans, for example?

Perhaps an even more telling criticism can be made. In a college or university, performance below a certain level means a bitter tragedy for a student. In an employment situation, job failure can often be equally damaging to self-esteem. In the selection situation described above, the rate of subsequent failure after admission would be one-fourth if the top half were admitted, but one-third if a Thorndikean admission rule were used. Furthermore most of the increase in failures comes precisely from the poor-risk admissions. Their failure rate is two-thirds. Thus in the end, a Thorndikean rule may be even more unfair to those at the bottom of the test score distribution than to those at the top.

Darlington's Definition No.3

Let us first review the introductory comments and analyses in Darlington's (1971) article and then consider his Definition No.3 in some detail. Darlington's first step was a restatement of the Cleary (1968) and Thorndike (1971) criteria for a "culturally fair" test in terms of correlation rather than regression. Again considering a comparison of blacks and whites, let X be the predictor, Y the criterion, and C the indicator variable for "culture" (i.e., C=1 for whites, C=0 for blacks). Darlington made the empirically plausible assumption that the groups have equal standard deviations on both predictor and criterion and that the validity of the predictor is the same (Schmidt, Berner, and Hunter, 1973) for both groups (hence parallel regression lines). Darlington then correctly noted

that Cleary's (1968) criterion for a "fair" test can be stated

$$r_{CY \cdot X} = 0$$

That is, there is no criterion difference between the races beyond that produced by their difference on X (if any). If all people are selected using a single regression line, then Thorndikean quotas are guaranteed by Darlington's Definition No.2, i.e.,

$$r_{CX} = r_{CY}$$

That is, the racial difference on the predictor must equal the racial difference on the criterion in standard score units. However, if people are selected using multiple regression or separate regression lines then this equation is not correct. Instead there are two alternate conditions:

$$R_{Y \cdot CX} = 1$$

or

$$r_{CY} = 0$$

That is, if separate regression lines are used, then the percentages selected match Thorndike's quotas only if the test has perfect validity or if there are no differences between the groups on the criterion.[5]

Darlington then attacked the Cleary definition on two bases: (1) the reliability issue raised by Linn and Werts (1971), which was discussed above, and (2) the contention that race itself would be a fair test by Cleary's definition. Actually if race were taken as the "test," then there would be no within-group variance on that predictor, and hence no regression lines to compare. Thus Cleary's definition cannot be applied to the case where race itself is used as the predictor test.[6] The nontrivial equivalent of this is a test whose sole contribution to predicting Y is the race differences on the mean of X. But for such a test, the regression lines are perfectly horizontal and grossly discrepant. That is, in a real situation, Cleary's definition would rule that a purely racial test is "biased."

Darlington's Definition No.3 and Cole's argument. Darlington (1971) proposed a third definition of test fairness, his Definition No.3. This definition did not attract a great deal of attention

until Cole (1973) offered a persuasive argument in its favor. We will first present Darlington's third definition, his justification of it, and our critique of that justification. We will then consider Cole's argument.

If X is the test and Y is the criterion, and C, the variable of "culture," is scored 0 for blacks, 1 for whites, then Darlington's Definition no.3 can be written as follows: The test is fair if

$$r_{XC \cdot Y} = 0.$$

His argument for this definition went as follows: The ability to perform well on the criterion is a composite of many abilities, as is the ability to do well on the test. If the partial correlation between test and race with the criterion partialed out is not zero, there is a larger difference between the races on the test than would be predicted by their difference on the criterion. Hence the test must be tapping abilities which are not relevant to the criterion but on which there are racial differences. Therefore, the test is discriminatory.

Note that Darlington's argument makes use of assumptions about causal inference. If those assumptions about causality are in fact false, then his interpretation of the meaning of the partial correlation is no longer valid. Are his assumptions so plausible that they need not be backed up with evidence? Consider the time-ordering of his argument. He is partialing the criterion from the predictor. In the case of college admissions, this means that he is calculating the correlation between race and entrance exam score with GPA four years later being held constant. This is looking at the causal influence of the future on the past and is only valid in the context of very special theoretical assumptions. The definition would in fact be inappropriate even in the context of a concurrent validation study, since concurrent validities are typically derived only as convenient estimates of predictive validity. Thus even when there is no time lag between predictor and criterion measurement, one is operating implicitly within the predictive validity model.

Let us explore the matter of causality more fully through the use of a concrete example. Let us consider a professional football coach attempting to evaluate the rookies who have joined the team as a result of the college "draft." Since the players have all come from different schools, there are great differences

in the kind and quality of training that they have had. Therefore the coach cannot simply rely on how well they play their positions at present; they will undergo considerable change as they learn the ropes over the next few months. What the coach would like to know is exactly what their athletic ability is, without reference to how well they have learned to play to date. Suppose he decides to rely solely on the 40-yard dash as an indicator of football ability, i.e., as a selection test. It is possible that he will then find that he is selecting a much larger percentage of blacks than he had using his judgment of current performance. Does this mean that the test discriminates unfairly against whites? That depends on the explanation for this outcome. Consider what is required of the defensive lineman on a passing play. His ability to reach the quarterback before the ball is thrown depends not only on the speed necessary to go around the offensive lineman opposing him, but also on his possessing sufficient arm strength to throw the offensive lineman to one side (defensive linemen can use their hands). Assume, for the sake of this example, that blacks are faster, on the average, than whites, but that there are no racial differences in upper body strength. Since the 40-yard dash represents only speed, and makes no measure of upper body strength, it cannot meet the requirements of Darlington's definition. That is, the 40-yard dash taps only the abilities on which there are racial differences and does not assess those which show no such differences.

How does the 40-yard dash behave statistically? If speed and upper body strength were the only factors in football ability and if performance on the 40-yard dash were a perfect index of speed, then the correlations would satisfy $r_{YC \cdot X} = 0$. That is, by Cleary's definition, the 40-yard dash would be an unbiased test. Since $r_{YC \cdot X} = 0$, $r_{XC \cdot Y}$ cannot be zero and hence, according to Darlington's definition, the 40-yard dash is "culturally unfair," i.e., biased against whites. (Since the number of whites selected would be fewer than the Thorndikean quota, Thorndike too would call the test biased.) If the coach were aware that upper body strength was a key variable and were deliberately avoiding the use of a measure of upper body strength in a multiple regression equation, then the charge that the coach was deliberately selecting blacks would seem quite reasonable. But suppose that the nature of the missing predictor (i.e., upper body strength) was completely unknown. Would it then be fair to charge the coach with using an unfair test?

At this point we should note a related issue raised by Linn and Werts (1971). They too considered the case in which the criterion is affected by more than one ability, one of which is not assessed by the test. If the test assessed only verbal ability, and the only racial differences were on verbal ability, then the situation would be like that described in the preceding paragraph: the test would be fair by the Cleary definition but unfair according to Darlington's definition No.3. However, if there are also racial differences on the unmeasured ability, then the test will not be fair by Cleary's definition. For example, if blacks were also lower, on the average, in numerical ability, and numerical ability was not assessed by the entrance test, then the black regression line and the test would be unfair to whites by Cleary's definition. According to Darlington's definition No.3, on the other hand, the verbal ability test would be fair if, and only if, the racial difference on the numerical test were of exactly the same magnitude in standard score units as the difference on the verbal test. If the difference on the missing ability were less than the difference on the observed ability, then Darlington's definition would label the test unfair to blacks, while if the difference on the missing ability were larger than the difference on the observed ability then the test would be unfair to whites. Furthermore, if the two abilities being considered were not the only causal factors in the determination of the criterion (e.g., if personality or financial difficulties were also correlated), then these statements would no longer hold. Rather, the fairness of the ability test under consideration would depend not only on the size of racial differences on the unknown ability, but on the size of racial differences on the other unknown causal factors as well. That is, according to Darlington's definition No.3, the fairness of a test cannot be related to the causal determinants of the criterion until a perfect multiple regression equation on known predictors has been achieved. Therefore, Darlington's definition can be statistically but not substantively evaluated in real situations.

For purposes of illustration, we now consider a simplified theory of academic achievement in college. Suppose that the college entrance test were in fact a perfect measure of academic ability for high school seniors. Why is the validity not perfect? Consider three men of average ability: Sam meets and marries Wonder Woman. She scrubs the floor, earns 200 dollars a week, and worships the ground Sam walks on. Sam carries a B average. Bill dates from time to time, gets hurt a little, turns off once or twice

statistical definition thus does not fit his substantive assumptions in this context--unless one is willing to accept luck as an "ability" and treat it as any other ability would be treated.

The problem with Darlington's definition becomes even clearer if we alter slightly the example in the above paragraph. Suppose that the world became more benign and that the tendency for blacks to have bad luck disappeared. Then, making the same assumptions as above (i.e., a perfect test and our theory of academic achievement), the regression curves would be equal, and $r_{YC \cdot X} = 0$. Thus according to Cleary's definition, the test would be unbiased against either group. Darlington's definition No.3 would now label the test as unfair to blacks. This last statement is particularly interesting. In our theory of achievement we have assumed that exactly the same ability lies at the base of performance on both the test and later GPA. Yet it is not true in our example that $r_{XC \cdot Y} = 0$. Thus this example has shown that Darlington's substantive interpretation of $r_{XC \cdot Y}$ does not hold with our additional assumption (of a non-statistical nature), and hence his argument as to the substantive justification of his definition is not logically valid.

We note in passing that our modified example poses a problem for Cleary's definition as well as for Darlington's. If the difference between the regression lines were in fact produced by group differences in luck, then would it be proper to label the test as biased? And if this model were correct, how many unqualified individualists would feel comfortable in using separate regression lines so as to take into account the fact that blacks have a tougher life (on the average) and hence make poorer GPA's, ability constant? In the case of both definitions, this analysis points up the necessity of substantive models and considerations. Statistical analyses alone can obscure as much as they illuminate.

As mentioned earlier, Darlington's definition No.3 received little attention until a novel and persuasive argument in its favor was advanced by Cole (1973). Her argument was this: Consider those applicants who would be "successful" if selected. Should not such individuals have equal probability of being selected regardless of racial or ethnic group membership? Under the assumption of equal slopes and standard deviations for the two groups, the answer to her question is in the affirmative only if the two

on girls who like him, and generally has the average number of ups and downs. Bill carries a C average. Joe meets and marries Wanda the Witch. She lies around the house, continually nags Joe about money, and continually tells him that he is sexually inadequate. As Joe spends more and more time at the local bar, his grades drop to a D average and he is eventually thrown out of school. In a nutshell, the theory of academic achievement that we wish to consider is this: achievement consists of ability plus luck, where luck is a composite of money troubles, sexual problems, automobile accidents, deaths in the family, and other incidents in personal history. Luck in this sense is a random variable but cannot be considered random error, since its effects are stable over time. According to this theory, a difference between the black and white regression lines (over and above the effect of test unreliability) indicates that blacks are more likely to have bad luck than whites are. Before going on to statistical questions, we note that because we have assumed a perfect ability test, there can be no missing ability in the following discussion. And because we have assumed that non-ability differences are solely determined by luck, the entity referred to as "motivation" is in this model simply the concrete expression of luck in terms of overt behavior. That is, in the present example, motivation is assumed to be wholly determined by luck and hence already included in the regression equation.

Now let us consider the statistical interpretations of the fairness of our hypothetical perfectly valid (with respect to ability) and perfectly reliable test. Since blacks are assumed to be unlucky, as well as lower, on the average, in measured academic ability, the racial difference in college achievement in this model will be greater than that predicted by ability alone and hence the regression lines of college performance onto ability will not be equal. Thus according to Cleary the test is biased against whites. According to Thorndike the test is probably approximately fair (perhaps slightly biased against blacks). According to Darlington, the test could be either fair or unfair. If the racial difference on luck were about the same in magnitude as the racial difference on the ability test, then the test would be fair. But if the racial difference on luck were less than the difference on ability, then the test would be unfair to blacks. That is, the Darlington assessment of the fairness of the test would not depend on the validity of the test in assessing ability, but on the relative harshness of the personal-economic factors determining the amount of luck accorded the two groups. Darlington's

regression lines of test on criterion are the same (and hence $r_{XC \cdot Y} = 0$). That is, Cole's definition is the same as Cleary's

with the roles of the predictor and criterion reversed. However, this similarity of statement does not imply compatibility--just the reverse. If there are differences between the races on either test or criterion, then the two definitions are compatible only if the test validity is perfect. So the two definitions are almost invariably in conflict.

Although Cole's argument sounds reasonable and has a great deal of intuitive appeal, it is flawed by a hidden assumption. Her definition assumes that differences between groups in probability of acceptance given later success if selected are due to discrimination based on group membership. Suppose that the two regression lines of criterion performance as a function of the test are equal (i.e., the test is fair by Cleary's definition). If a black who would have been successful is rejected while a white who fails is accepted, this need not imply discrimination. The black would not be rejected because he is black, but because he made a low score on the ability test. That is, the black would have been rejected because his ability at the time of the predictor test was indistinguishable from that of a group of other people (of both races) who, on the average, would have had low scores on the criterion.

To make this point more strongly, we note that according to Cole's definition of a fair test, it is unethical to use a test of less than perfect validity. To illustrate this, consider the use of a valid ability test to predict academic achievement in any one group, say whites, applying for university admission. If the university decides to take only the people in the top half of the distribution of test scores, then parents of applicants in the bottom half, acting under Cole's definition, might well file suit charging discriminatory practice. According to Cole, an applicant who would be successful if selected should have the same probability of being selected regardless of group membership. That is, among the applicants who would have been successful had they been selected, there are two groups. One group has a probability of selection of 1.00 because their score on the entrance exam is higher than the cut-off. The other group of potentially successful applicants has a selection probability of .00 because their exam score is lower than the cut-off. According to Cole, we should ask: "Why should a person who would be successful be denied a college berth merely because he had a low test score? After all it's success

that counts, not test scores." But the fact is that, for any statistical procedure that does not have perfect validity, there must always be applicants who will be incorrectly predicted to have low performance, i.e., there will always be successful people whose predictor score is down with the generally unsuccessful people instead of up with the generally successful people (and vice versa). In that sense, anything less than a perfect test will always be "unfair" to the potentially high achieving people who were overlooked. It can be seen that lack of perfect validity functions in exactly the same way as test unreliability, discussed earlier.

As noted earlier in the case of Thorndike's definition, this problem could be partly overcome in practice if restrictions arrived at by social consensus could be put on the defining of "bona fide minority groups." But given the almost unlimited number of potentially definable social groups, it is unlikely that social or legal consensus could be reached limiting the application of this definition to blacks, Chicanos, American Indians, and a few other groups.

Basically Cole has noted the same fact that Thorndike noted: that in order for a test with less than perfect validity to be "fair" to individuals, the test must be "unfair" to groups. In particular, in our example, the group of applicants who score below average on the test will have none of their members selected despite the fact that some of them would have shown successful performance if selected. It is thus "unfair" to this group. However, it is "fair" to each individual, since each is selected or rejected based on the best possible estimate of his future performance. It is perhaps important to note that this is not a problem produced by the use of psychological tests; it is a problem inherent in selection decisions. Society and its institutions must make selection decisions. They are unavoidable. Elimination of valid psychological tests will usually mean their replacement with devices or methods having <u>less</u> validity (e.g., the interview), thus further increasing the "unfairness" to individuals or groups.

<u>Darlington's Definition No.4</u>

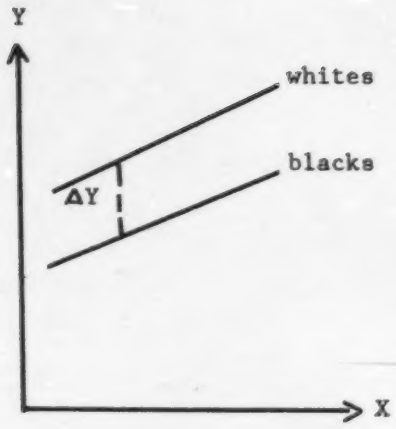
The fourth concept of test bias discussed by Darlington (1971) defines a test as fair only if $r_{CX} = 0$. Hence, any test which shows any ethnic differences at all in mean score is considered unfair, regardless of the magnitude of the group difference in

performance. This concept of test fairness corresponds directly to the ethic of population-proportion-based quotas.

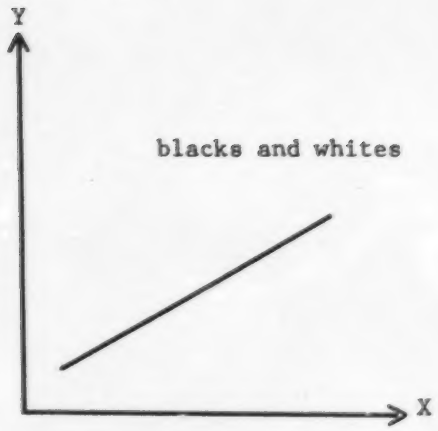
A Fifth Definition of Test Fairness

After defining and discussing four different statistical models of test fairness, Darlington (1971) turned to the commonly occurring prediction situation in which there is a difference favoring whites on both the test and the criterion and the black regression equation falls below that for whites. This situation is shown in Figure 3a. Noting that the use of separate regression equations (or the equivalent, use of a multiple regression equation with race as a predictor), as required by Cleary's (1968) definition, would often admit or select only an extremely small percentage of blacks, Darlington introduced his concept of the "culturally optimal" test. Under this concept, admissions officers at a university, for example, are asked to consider two potential graduating seniors, one white and the other black, and to indicate how much higher the white's GPA would have to be before the two candidates would be equally attractive. This number is symbolized K and given a verbal label such as "racial adjustment coefficient." Then in determining the fairness of the test, K is first subtracted from the actual criterion scores (GPA's) of each of the white subjects. If these altered data satisfy Cleary's (1968) definition of a fair test, the test is considered "culturally optimal."

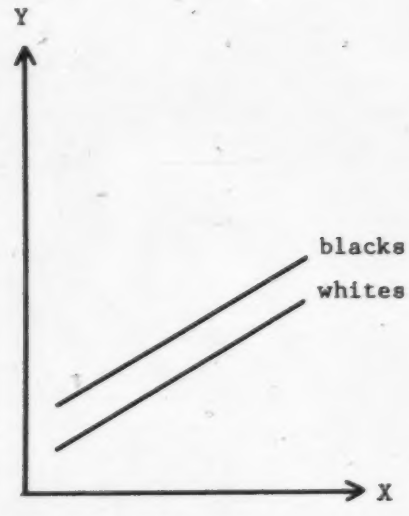
Figure 3 illustrates the geometrical meaning of Darlington's altered criterion. If the admissions officer chooses a value of K which is equal to Y in Figure 3a, then the altered data will appear as in Figure 3b, i.e., there will be a single common regression line and the test as it stands will be "culturally optimal." If, however, an overzealous admissions officer chooses a value of K greater than Y, then the altered data will appear as in Figure 3c, i.e., the test will be biased against blacks according to Cleary's (1968) definition and will thus not be "culturally optimal." Although Darlington is willing to tamper with criterion scores, he does not allow for application of this process to predictor scores. Thus if the situation shown in Figure 3b obtains, it can be corrected only by (1) modifying the factor structure of the test in such a way that the data move to the configuration in Figure 3b or (2) abandoning the test and seeking another which meets the requirements of Figure 3b. Similarly, should an uncooperative admissions officer select $K < \Delta Y$, and thus produce the situation shown in Figure 3d, the only remedies are changes



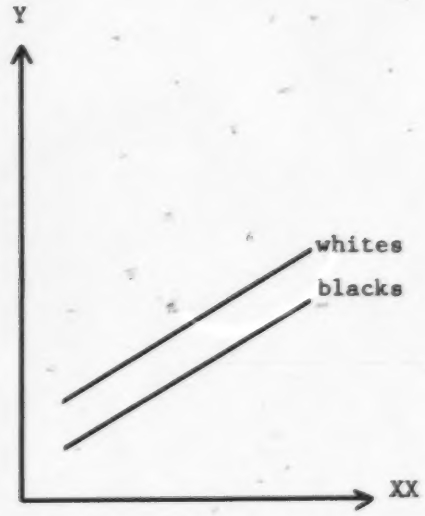
3a. The original data



3b. The altered data when K = ΔY



3c. The altered data wnen K > ΔY



3d. The altered data when K < ΔY

Figure 3. Darlington's (1971) method of altering the data to define a "culturally optimal" test.

in the nature of the test or the introduction of an entirely new test.

What is the end result of these complicated procedures? From a statistical point of view, subtracting a constant from the criterion scores of whites is identical in its results to adding an equivalent constant to black predictor scores without changing the black prediction equation. Either procedure can be used to alter the prediction situation so as to create the impression of a single regression line when in fact group intercepts are different. If a single regression line is then used in practice, race as a predictor of performance is ruled out. Thus this definition of Darlington's corresponds, perhaps ironically, to qualified individualism.

II. ETHICAL POSITIONS, STATISTICAL DEFINITIONS, AND PROBLEMS

Frank L. Schmidt
Personnel Research and Development Center
U.S. Civil Service Commission

In this section, we briefly relate each ethical position to its appropriate statistical operationalizations and point out some of the advantages and disadvantages of each ethical approach.

Unqualified Individualism: Models and Problems

The remedies advanced for tests unfair by Cleary's (1968) definition make clear that the Cleary approach is one of unqualified individualism. The recommended use of separate regression equations when these are not equal for all groups is clearly equivalent to the use of race as a predictor of performance. It is this requirement that race, sex, age, and other such predictors must be employed when valid that would seem to create legal problems for unqualified individualism. The 1964 Civil Rights Act, the 1972 Equal Employment Opportunity Act, and other such legislation specifically forbid personnel decision-making based on these variables. Ironically, the Cleary (1968) approach to test fairness, with its required adjustments, is endorsed, recommended, and even required by the guidelines on employment testing published by both the U.S. Equal Employment Opportunity Commission (1970) and

the Office of Federal Contract Compliance (1971). These guidelines require that, where valid, race be used as a predictor of job success, even though the laws and executive orders on which the guidelines are ostensibly based seem to be clear in forbidding such use. There is no ready explanation for this apparent contradiction.

There is another ethical imperative inherent in unqualified individualism which gives rise to potential problems--in this case technical, rather than legal problems. Unqualified individualism requires that one apply to each candidate that prediction procedure that is most valid for him. Although it is technically impossible to develop separate procedures for each individual, it is often feasible to develop separate regression equations for different groups, and in these cases the ethics of unqualified individualism requires that this be done. This could theoretically lead to the impossible task of constructing different prediction equations for all definable subpopulations. It is also important to note that each such equation must be maximally valid, because low predictability within any group, relative to other groups, leads to greatly reduced selection opportunities for members of that group, at least at the selection ratios commonly in use(i.e., SR $<$ .50).

Consider, for example, the situation in which validity is zero for one group. The predicted criterion score for everyone in that group is then the same: the mean criterion score for that group. Thus either everyone in that group is accepted or everyone in the group is rejected. If that group is in fact highly homogeneous on the criterion, then this is perfectly reasonable. But if the zero-validity group has the same degree of spread on the criterion as other groups, then this lack of discrimination poses ethical problems: either a great many poor prospects are being admitted, or a great many excellent prospects are being overlooked.

Fortunately for the ethic of unqualified individualism, the research evidence strongly indicates that differential validity by race is no more than a chance phenomenon (Schmidt, Berner, and Hunter, 1973). The same may later be shown with respect to other population subgroups, thus greatly reducing the scope of this problem. The problem would not thus be eliminated, however; although the same tests may be valid across population subgroups, and regression slopes may be equal, there is much research evidence

(Ruch, 1972; Stanley, 1971; Temp, 1971; Campbell, et al., 1973) that intercepts often differ significantly. Thus the task of testing and adjusting for intercept differences remains.

Qualified Individualism: Models and Problems

Of the statistical models of test fairness we have reviewed, only Darlington's fifth definition, his "culturally optimal" test, corresponds exactly to the ethic of qualified individualism. It should be added that, although this conclusion is certain, it is difficult to be certain in reading Darlington's (1971) article that it is, in fact, what he initially intended. It should be noted that this ethical position does not require that variables like race, sex, religion, etc., not be statistically valid predictors of future performance, but only that, should they be valid, they must not be used. Thus both tests that meet, and those that do not meet, Cleary's (1968) requirements are defined as "fair" so long as a single regression is used, and, of course, the complicated analysis recommended by Darlington (1971) to assure "cultural optimality" is not required in practice. This position seems to come closest to that embodied in the various civil rights laws, and, as such, might be expected to encounter few legal hazards. However, as noted above, EEOC and OFCC employment testing guidelines have apparently endorsed the Cleary (1968) approach to test fairness and thus have adopted a position of unqualified individualism. This apparent contradiction between the wording and intent of the law and the Federal rules designed to enforce the law is unexplained. It does, however, raise the possibility - however remote - that employers and institutions adopting a position of qualified individualism could have charges of discrimination brought against them.

Advocates of qualified individualism must face another, more subtle but perhaps more real, problem. The ethical imperative here requires that the prediction equation that has maximum validity for the entire population - without regard to group membership - be identified and employed. But there is a difficulty in doing this. Suppose, for example, that for a certain city college the black regression line falls below the white regression line, i.e., race is a valid predictor for that college. Although use of race as a predictor is, of course, forbidden to the qualified individualist, there may be alternative ways of increasing

the overall validity of the prediction equation that are equally objectionable. For example, if race is a valid predictor, then a properly coded version of the student's address may also be a valid predictor and increase overall validity. This indirect indicator of race would probably be detected and rejected, but a more subtle cue might not be properly identified. The most subtle problem is the one facing the test constructor: if the black regression line falls below the white regression line, then the introduction of items whose content is biased against blacks would increase the overall validity of the test. If the separate regression lines of the unqualified individualist are used, then racially biased test material would have no effect on the selection of applicants. But if that is forbidden, then material biased against blacks would lower the black scores on the predictor and hence make their scores using the white regression line more accurate. That is, the introduction of material biased against blacks would reduce the overprediction of black performance and hence might raise the validity of a single regression line use of the test.[7]

The problem in its general form, then, is that any measured variable which correlates with race, sex, religion, etc. (i.e., shows group differences) can be considered to be an indirect (and imperfect) indicator of group membership. Since he is forbidden to use group membership itself as a predictor even if valid, the qualified individualist may be tempted to substitute indirect indicators of group membership that may be "unfair." How can he decide whether a given race-correlated predictor is "fair" or "unfair"? This question can be answered, although all answers involve some element of judgment. The first criterion on which this judgment can be based is the apparent "intrinsicalness" of the relationship between the predictor and performance. If the predictor is a job sample test (e.g., a typing test) assessing the skills actually required on the job, there is little doubt that the relation is intrinsic. Scores on a written achievement test could also easily pass this test, as would a face-valid aptitude test. Scores on a weighted biographical information inventory, on the other hand, would be allowed only if they were able to meet the second, less subjective standard: validity coefficients large enough to be practically significant for both groups separately. A predictor which is valid only by virtue of its correlation with race will show no within-group validity.

Thus the qualified individualist's answer to the question posed in the example above is that if the material to be added

to the test appears to have an intrinsic relation to performance and if it increases both within-group validities, it is ethically admissible. It is not "biased" against blacks as blacks but merely against applicants (of whatever race) who are less capable of performing well on the criterion. The fact that the proportion in the low skill group is greater for blacks than for whites is an ethically irrelevant fact. If it is ethically permissible to employ a test which shows racial differences, then it must follow that it is ethically permissible to improve the validity of the test within the rules described above. In fact, many qualified individualists would probably require only that the added material meet one of the two criteria. If the material is intrinsically related to job performance, one need not demonstrate significant within-group validities, and vice versa.

While the preceding distinctions are regarded as crucial among qualified individualists, they receive short shrift from those committed to other ethical positions. For one whose ethical position is that of population-based quotas, any predictor or test material showing racial differences is ipso facto unfair. The correlation between test and race is greater than zero and its use will produce selection ratios different from population percentages. To the unqualified individualist, any variable which shows a statistically reliable correlation with performance-- regardless of within-group validities or content --is fair and there is a positive ethical obligation to employ it.

The Quota Ethic: Models and Problems

Only Darlington's definition No. 4, which requires the complete absence of racial or ethnic differences (i.e., $r_{CX} = 0$) corresponds to the ethical position of population-based quotas. But quotas can be set on bases other than population percentages. Darlington's definition No. 3 offers one such basis, and Thorndike's (1971) model of test fairness (i.e., $r_{CX} = r_{CY}$) sets selection quotas for population subgroups based on past performance of each group as a whole. Thus, the individual inclined toward a quota ethic may choose any of these definitions, depending on how far he chooses to carry the quota concept; Thorndike's (1971) model represents the smallest departure from the concept of individual merit, and population-based quotas, the greatest. [8]

As with unqualified individualism, the ethic of quotas is potentially susceptible to legal problems. It is based on the legally uncertain proposition that ethnic and social groups as such, as well as individuals, have constitutional and legal rights. Our legal and governmental system, on the other hand, is largely built around the idea of individual rights. If only individuals have rights, then all quota-based systems, in varying degree, are unconstitutional, since they require that decisions on the basis of individual qualities and qualifications must be sacrificed to the attainment of the proper group ratios--which, in turn, are based on the idea of group, rather than individual, rights. (Ironically, quota-based systems may be illegal for the same fundamental reason that unqualified individualism is: both ethical system require decisions to be made, at least to some extent, directly on the basis of group membership.) A recent case which would perhaps have done much to clarify this issue was sidestepped by the U.S. Supreme Court (Defunis vs. Odegaard). But there are many such "reverse discrimination" suits pending in the courts, and the legal issue will almost certainly be addressed by the Supreme Court in the future.

The second major problem characterizing quota-based ethical systems is that the criterion performance of selectees as a whole can be expected to be considerably lower than under unqualified or even qualified individualism (Hunter, Schmidt, and Raushenberger, in press). In college selection, for example, the poor-risk blacks who are admitted by a quota are much more likely to fail. Thus in situations where low criterion performance carries a considerable penalty, being selected on the basis of quotas is a mixed blessing. Second, there is the effect on the institution. The greater the divergence between the quotas and the selection percentages based on actual expected performance, the greater the difference in mean performance in those selected. If lowered performance is met by increased rates of expulsion or firing, the quotas are undone and there is considerable anguish for those selected who did not succeed. Furthermore, the public image of the institution may suffer as a result of the high rate of expulsion. On the other hand, if the institution tries to adjust to the candidates selected, there may be great cost and inefficiency (Hunter, Schmidt, and Raushenberger, in press). In the case of academic institutions, quotas inevitably lower the average performance of graduates and hence the prestige rating of the school. Similar considerations apply in the case of the employment setting, but here the direct and immediate impact on individual welfare is often greater. For example,

assuming valid selection tests and other instruments, air traffic controllers hired under any of the quota systems rather than under the Cleary model would be more likely to make the kinds of errors that can lead to air disasters. Truck drivers selected under a quota system would be more likely to be involved in accidents on the road. Thus in the employment setting differences between the various models of fairness often translate not only into economic loss but also into the most precious of all commodities, human lives.

A final, and less momentous, consideration in the case of quota-based ethical systems concerns methods of selection to be used within groups once group quotas have been set. Most advocates of quota-based systems would probably involve individualism at this point, selecting those within each group with the highest predicted performance. This resort to individualism within groups, rather than random selection, mitigates somewhat the negative impact of the quota system on selectee performance. It also makes clear the underlying ethical assumptions of this approach: (1) ethnic groups per se have legal rights and these rights override those of individuals where there is a conflict, and (2) the individual's right to be considered on the basis of his qualifications should be recognized when it does not conflict with group rights (i.e., within ethnic groups).

Ethical Systems, Statistical Models, Individual Merit, and Social Goals

The ethical systems and statistical models of decision fairness reviewed in this paper may be scaled along a dimension that might be called "emphasis on individual merit." The systems and models at the high end of this continuum are based on the assumption that the right of the individual to be considered on the basis of his qualifications and expected performance is paramount. Those at the low end assume that the rights of groups, and social goals and considerations in general, take precedence over individual rights whenever there is a conflict. The ordering of the models and systems along this continuum is as follows: (1) Cleary's (1968) approach, corresponding to unqualified individualism; (2) Darlington's (1971) "culturally optimal" test (i.e., his fifth definition), corresponding to qualified individualism; (3) Thorndike's (1971) limited quota model; (4) the more extreme quota model represented by

Darlington's (1971) definition No.3 and Cole's (1973) definition; and (5) Darlington's (1971) definition No.4, corresponding to a population-based quota system. Selection tests currently used in employment and education tend to fall somewhere between the Cleary and Thorndike models (Schmidt and Hunter, 1974; Linn, 1973; Campbell, et al., 1973), that is, in the general region of qualified individualism. Unqualified individualists must conclude that tests are often slightly biased against the majority group, while to Thorndikeans they are somewhat unfair to the minority group. Those who adhere to Darlington's definition No.3 or to the ethic of population-based quotas must feel that current tests are markedly unfair to minority groups. The qualified individualist, of course, concludes that most currently used tests are probably reasonably close to being fair to all groups.

Which of these definitions will ultimately prove most acceptable - legally, socially, and ethically - to the American people? The answer is not yet known or knowable, but it is certain to depend on at least three important factors: (1) the strength of the commitment of the general public to the idea of individual merit, (2) public support of the national commitment to increased minority income, educational, and occupational levels, and (3) perhaps most importantly, the coming court rulings on the delicate issue of individual rights versus group rights and social goals.[9] Under these circumstances, and given the inherent subjectivity of decisions in this area, it would be highly inappropriate for us to urge any one of these ethical positions or statistical models on psychology as a whole. But it is our hope that, by explicating the important differences among the various options, this paper will contribute to the making of informed, intelligent decisions.

APPENDIX

This appendix contains the mathematical calculation of the expected achievement level of the group that would be selected by the full application of Thorndike's criterion, i.e., a group selected so that for each test score x, the number of people selected is proportional to the probability that persons at that test level would in fact be "successful". The definition of "successful" used below is "performance above average on the criterion". That is, the calculations done below assume a base rate of .50. The selection ratio assumed is also .50.

For simplicity, both test and performance have been assumed to be measured in standard scores. The symbol $\phi(x)$ is the standard normal density function and the symbol $\Phi(x)$ is the standard normal cumulative distribution function. The symbol A will be used for "accepted" (or selected for admission).

If the criterion of success is the top 50 percent, then in terms of standard scores, the success criterion is $Y > 0$. Thus the conditional probability of being accepted is

$$P(A|X) = P\{Y>0|X\}$$

$$= P\left\{ \frac{y-rx}{\sqrt{1-r^2}} > -\frac{rx}{\sqrt{1-r^2}} \right\}$$

$$= 1 - \Phi\left(-\frac{rx}{\sqrt{1-r^2}}\right)$$

$$= \Phi\left(\frac{r}{\sqrt{1-r^2}}\, x\right)$$

If the number selected at each test score is $P(A|X)$, then the overall selection ratio will be

$$P(A) = \int \Phi\left(\frac{r}{\sqrt{1-r^2}}\, x\right) \phi(x)\, dx = \frac{1}{2}$$

The distribution of the test score among those selected is

$$f_A(x) = \frac{P(A|X)\ \ P(X)}{P(A)} = 2\ \Phi\left(\frac{r}{\sqrt{1-r^2}}\ x\right)\ \phi(x)$$

Since X and Y are in standard score form, the regression of Y on X is given by

$$E(Y|X) = rx$$

Thus the mean criterion score among those selected will be

$$E(Y) = E(E\{Y|X\})$$

$$= \int rx\ f_A(x)\ dx$$

$$= \int rx\ \ 2\Phi\left(\frac{r}{\sqrt{1-r^2}}\ x\right)\ \phi(x)\ dx$$

This is not an easy integral to calculate, and the calculation below will thus be broken into five steps. First to simplify the algebra, we will introduce the parameter α by the definition

$$\alpha = \frac{r}{\sqrt{1-r^2}}$$

In particular, if r = .6 then

$$\alpha = \frac{.6}{\sqrt{1-.36}} \quad \frac{.6}{.8} \quad \frac{3}{4}$$

The formula for mean criterion performance among those selected can then be written

$$E(Y) = \int 2r\ \Phi(\alpha x)\ \phi(x)\ dx$$

Step 1 First we apply the method of integration by parts:

$$\int x\ \Phi(\alpha x)\ \phi(x)\ dx = \int \Phi(\alpha x)\ \{x\phi(x)\ \}\ dx$$

$$= \Phi(\alpha x)\ \ u(x) - \int u(x)\ \{\ \Phi'(\alpha x)\ \alpha\ \}\ dx$$

where

$$u(x) = \int x\phi(x)\, dx = -\frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Step 2 Thus we have the definite integral:

$$\int_{-\infty}^{+\infty} x\, \Phi(\alpha x)\ \ \phi(x)\, dx = \Phi(\alpha x)\, u(x)\Big|_{-\infty}^{+\infty} - \int u(x)\ \ \phi'(\alpha x)\alpha\, dx$$

$$= 0 - \int u(x)\, \frac{e^{-\alpha^2 x^2/2}}{\sqrt{2\pi}}\, \alpha\, dx$$

$$= \frac{\alpha}{2\pi}\int e^{-x^2/2}\ \ e^{-\alpha^2 x^2/2}\, dx$$

$$= \frac{\alpha}{2\pi}\int e^{-(1+\alpha^2)x^2/2}\, dx$$

$$= \frac{\alpha}{2\pi}\int e^{-\beta^2 x^2/2}\, dx$$

where

$$\beta = \sqrt{1 + \alpha^2}$$

Step 3 Using the substitution $x = \frac{1}{\beta}y$, we can calculate the following integral:

$$\int e^{\frac{-\beta^2 x^2}{2}}\, dx = \int e^{\frac{-y^2}{2}}\, \frac{1}{\beta}\, dy$$

$$= \frac{1}{\beta}\sqrt{2\pi}$$

Step 4 Thus, we can finally calculate the main integral:

$$\int_{-\infty}^{+\infty} x\,\phi(\alpha x)\,\phi(x)\,dx = \frac{\alpha}{2\pi}\frac{1}{\beta}\sqrt{2\pi} = \frac{1}{\sqrt{2\pi}}\,\frac{\alpha}{\sqrt{1+\alpha^2}}$$

Since α was defined to be

$$\alpha = \frac{r}{\sqrt{1-r^2}}$$

we have

$$\frac{\alpha}{\sqrt{1+\alpha^2}} = \frac{r\sqrt{1-r^2}}{\sqrt{1+\frac{r^2}{1-r^2}}} = r$$

Step 5 Finally we can use the main integral to calculate the expected achievement level:

$$E(Y) = 2r\ \text{Integral} = 2r\,\frac{1}{\sqrt{2\pi}}\,\frac{\alpha}{\sqrt{1+\alpha^2}}$$

$$= \frac{2r^2}{\sqrt{2\pi}}$$

For r = .6, this formula yields E(Y) = .288

FOOTNOTES

1. This phenomenon would account for perhaps half of the magnitude of overprediction of black college grade point average found in the literature. In standard score units, the difference in intercepts due to unreliability is $\Delta y = (1-r_{xx})(\mu_W-\mu_B)$ where r_{xx} is the test reliability and $\mu_W - \mu_B$ is the white - black mean difference on the criterion (about one S.D.). For $r_{XX} = .80$, this would be only .2 S.D. whereas in the data reported in Linn (1973), the overprediction is about .37 S.D.

2. The reader may wonder why we show so much concern with the reliability of the test and no concern with the reliability of the criterion. Actually despite its large effect on the validity coefficient, no amount of unreliability in the criterion has any effect on the regression line of criterion on predictor. Let the true score equations for X and Y be $X = T + e_1$ and $Y = U + e_2$ and let the regression true score equation be $U = \alpha T + \beta$. Then the observed regression line will not have the same coefficients. Let the observed regression line be $Y = aX + b$. The slope of the observed regression line will be

$$a = r_{XY}\frac{\sigma_Y}{\sigma_X} = (r_{TU}\, r_{TX}\, r_{UY})\frac{\sigma_Y}{\sigma_X} = r_{TU}\frac{\sigma_T}{\sigma_X}\frac{\sigma_U}{\sigma_Y}\frac{\sigma_Y}{\sigma_X}$$

$$= r_{TU}\frac{\sigma_U}{\sigma_T}\frac{\sigma_T}{\sigma_X}\frac{\sigma_T}{\sigma_X} = \left(r_{TU}\frac{\sigma_U}{\sigma_T}\right)\left(\frac{\sigma_T^2}{\sigma_X^2}\right)$$

$$= \alpha\, r_{XX}$$

That is, the slope of the observed regression line is the slope of the true score regression line multiplied by the reliability of X. However, note that the slope of the observed regression line is completely independent of the reliability of Y. The intercept of the observed regression line is given by:

$$b = \mu_Y - \alpha\mu_X = \mu_U - \alpha\mu_T = \mu_U - r_{XX}\alpha\mu_T$$

Thus the intercept is also affected by the reliability of X, but is completely independent of the reliability of Y. If we have equal slopes on the true score regression equations and equal within-groups test reliability, any differences in the regression lines will be equal to

the difference between the intercepts and hence independents of r_{YY}. In the case where the true score regression lines are the same, the difference between the observed regression lines is

$$b_W - b_B = (1 - r_{XX}) (\mu_{UW} - \mu_{UB})$$

3. While on the topic of reliability, we should note that as the reliability approaches .00, the test becomes a random selection device and is hence utterly reprehensible to an individualist of either stripe. On the other hand a totally unreliable test would select blacks in proportion to population quotas. Ironically, the argument that tests are biased against blacks because they are unreliable is not only false, it is exactly opposite to the truth.

4. What Thorndike has rediscovered has long been known to biologists: Bayes' law is cruel. For example, if one of two equally reproductive species has a probability of .49 for survival to reproduce and the other species is .50, then ultimately the first species will be extinct. Maximization in probabilistic situations is usually much more extreme than most individuals expect (Edwards and Phillips, 1964).

5. Since the groups have equal standard deviations on both predictor and criterion, assume for algebraic simplicity that the variables have been scaled so that all within group standard deviations are unity. This means that deviation scores are standard scores. Suppose that the selection ratio for whites has been determined. Then there is a corresponding standard score on Y say Y^* such that the standard score $Y^* - \overline{Y}_W$ would cut off that percentage of whites. To select that same percentage of whites, there is a predictor score on the test, X_W, such that

$$X_W^* - \overline{X}_W = Y^* - \overline{Y}_W$$

If the multiple regression equation is

$$\hat{Y} = \alpha X + \beta C + \gamma$$

then the multiple regression cutoff score is

$$\hat{Y}^* = \alpha X_W^* + \beta + \gamma$$

$$= \alpha (X_W^* - \overline{X}_W + \overline{X}_W) + \beta + \gamma$$

$$= \alpha (X_W^* - \overline{X}_W) + \alpha \overline{X}_W + \beta + \gamma$$

Since multiple regression always matches the group means perfectly,

$$\bar{Y}_W = \alpha \bar{X}_W + \beta + \gamma$$

and hence

$$\hat{Y}^* = \alpha (X_W^* - \bar{X}_W) + \bar{Y}_W$$

The predictor cutoff score for blacks is determined by

$$\begin{aligned}\hat{Y}^* &= \alpha X_B^* + \gamma \\ &= \alpha (X_B^* - \bar{X}_B + \bar{X}_B) + \gamma \\ &= \alpha (X_B^* - \bar{X}_B) + \bar{X}_B + \gamma\end{aligned}$$

Since multiple regression matches means

$$\bar{Y}_B = \alpha \bar{X}_B + \gamma$$

and hence the black predictor cutoff satisfies

$$\hat{Y}^* = \alpha (X_B^* - \bar{X}_B) + \bar{Y}_B$$

Thorndike's quota for blacks is obtained if the standard score for the predictor cutoff is the same as the standard score for the criterion cutoff, i.e., if

$$X_B^* - \bar{X}_B = Y^* - \bar{Y}_B$$

Now we have in general

$$\alpha (X_B^* - \bar{X}_B) = Y^* - \bar{Y}_B$$

Thus Thorndike's quotas obtain only if

$$\begin{aligned}\alpha (Y^* - \bar{Y}_B) &= Y^* - \bar{Y}_B \\ &= \{\alpha (X_W^* - \bar{X}_W) + \bar{Y}_W - \bar{Y}_B \\ &= \alpha(Y^* - \bar{Y}_W) + \bar{Y}_W - \bar{Y}_B\end{aligned}$$

This is only true only if

$$\alpha(\bar{Y}_W - \bar{Y}_B) = \bar{Y}_W - \bar{Y}_B$$

Thus Thorndike's quotas are obtained only if one of two things is true: either = 1 or both sides are zero, i.e., either $r_{xy} = 1.00$ or $\overline{Y}_W - \overline{Y}_B = 0$.

Since the variables were all scaled to have equal within group standard deviations, the regression weight is in fact the within group predictor-criterion correlation. Thus = 1 means that the test has perfect validity.

The equation $\overline{Y}_W - \overline{Y}_B = 0$ is equivalent to $\overline{Y}_W = \overline{Y}_B$, i.e., no group difference on the criterion and hence $r_{CY} = 0$.

6. Darlington's error was a subtle one. He assumed that $r_{CY \cdot C} = 0$ when in fact $r_{CY \cdot C} =$ which is undefined.

7. This argument, of course, assumes that the newly added biased material has no detrimental effect on the within-group validities. We return to this consideration later.

8. Degree of departure from the concept of individual merit is directly related to loss of selection utility occasioned by use of the fairness model (Hunter, Schmidt, and Rausheuberger, in press).

9. On March 11, 1975, Federal Judge Spencer Williams, Unites States District Court, Northern District of California ruled in Cortez vs. Rosen that the Cleary model is the "only one which is historically, legally, and logically required". This ruling which sustained the use of a police examination shown to meet Cleary model requirements, is the first to address the question of the relative legal merits of alternative fairness models.

REFERENCES

Cleary, T.A. Test bias: prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cole, N.S. Bias in selection. Iowa City, Iowa: The American College Testing Program, 1972. (Also in Journal of Educational Measurement, 1973, 10, 237-255.)

Campbell, J.T., Crooks, L.A., Mahoney, M.H., and Rock, D.A. An investigation of sources of bias in the prediction of job performance: a six year study. Final Project report PR-73-37, Educational Testing Service, Princeton, NJ, 1973.

Darlington, R.B. Another look at "cultural fairness". Journal of Educational Measurement, 1971, 8, 71-82.

Supreme Court of the United States. Defunis v. Odegaard decision. Washington, D.C.: Author, March 1974.

Edwards, W. and Phillips, L.D. Man as transducer for probabilities in Bayesian command and control systems. In M.W. Shelley II and G.L. Bryan (eds) Human Judgments and Optimality. New York: Wiley, 1964.

Hunter, J.E., Schmidt, F.L., and Raushenberger, J. Fairness of psychological tests: implications of three definitions for selection utility and minority luring. Journal of Applied Psychology, in press.

Linn, R.L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Linn, R.L. and Werts, C.E. Considerations for studies of test bias. Journal of Educational Measurement, 1970, 7, 1-4.

Ruch, W.W. A re-analysis of published differential validity studies. Paper presented at the Symposium, Differential Validation Under EEOC and OFCC Testing and Selection Regulations.

Schmidt, F.L., Berner, J.G., and Hunter, J.E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 58, 5-9.

Schmidt, F.L., and Hunter, J.E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 1974, 29, 1-8.

Stanley, J.C. Predicting college success of the educationally disadvantaged. Science, March 19, 1971, 171, 640-647.

Temp, G. Validity of the SAT for blacks and whites in thirteen integrated institutions. Journal of Educational Measurement, 1971, 8, 245-251.

Thorndike, R.L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

U.S. Equal Employment Opportunity Commission. Guidelines on employment selection procedures. Washington, D.C.: Author, 1970.

U.S. Office of Federal Contract Compliance. Regulations on Employee Testing and Other Selection Procedures. U.S. Department of Labor, Washington, D.C.: Author, 1971.

GPO 908-754